Poster II-47

Phrase-Based Information Retrieval: Improvement Over the Word-Based System

Yang, Xiaohong[1], Zhou, Lucy[1], Lei, Xiaojun[2]*
LangPower Computing, Inc., [1]Tacoma, WA, and [2]Durham, NC, USA

Finding articles of interest in neuroscience has been a time-consuming process for scientists. General content management practice employs keyword-based indexing mechanism combined with statistical pattern analysis tools to retrieve relevant documents. Because words alone are not an appropriate representation of texts, the retrieval accuracy is low. In contrast, phrase-based indexes are more appropriate representation of texts. Therefore, this pilot study was aimed to test the hypothesis that phrase-based indexing would enhance the information retrieval (IR) accuracy employing a design that stores the syntactic structure of noun phrases. A set of 104 sentences from the neuroscience literature was selected and noun phrases were extracted according to their syntactic structure. These phrases were then parsed into pre-modifier of a head-noun, head-noun, and post-modifier of the head-noun and stored in a relational database. Query phrases were also parsed syntactically. The matching was performed in the following sequence: 1. match the head-noun of the query to head-nouns in the dataset; 2. match the pre-modifiers of positive matches of head-nouns; 3. match the post-modifier. The search result was compared with the conventional word-based searches such as "and," "or," and the whole phrase. The precision and recall of the search results were calculated and student t analysis was used for the statistical analysis. Results: The dataset consists of 2,184 words and 14,365 characters. These sentences were parsed into 581 noun phrases with 504 pre-modifiers, 564 head-nouns, and 195 post-modifiers. The results showed that there is a significant increase in the precision of searches using phrase-based search algorithm comparing to word-based "and," "or," or whole phrase searches in certain categories (see table 1). This pilot study is the first step in comparing phrase-based indexing and keyword-based indexing in IR. Its result will be extended to IR at an article level. Conclusion: Applying phrase-based indexing to the neuroscience literature offers improved search accuracy comparing to the conventional word-based indexing methods. The study suggests that phrase-based indexing has the potential to aid biomedical researchers to efficiently manage their research data and literature.

Table 1. Comparison of phrase-based and word-based searches.

Query		2-component query[#] (n=26)	3-componet query[##] (n=19)
Phrase-based match	Precision	95.8±3.8%	84.2±8.6%
	Recall	86.3±8.3%	84.2±7.4%
	Number of return	1.8±0.5	0.94±0.12
Word-based "and" match	Precision	74.0±8.5% **	88.6±6%
	Recall	91±6%	94.7±5.3%
	Number of return	4.2 ± 1.4	1.21±0.16
Word-based "or" match	Precision	13.7±4.3% ***	12.6± 3.2% **
	Recall	94±5.2%	100±0%
	Number of return	31.6±8.1 ***	15.7±2.8 ***
Whole phrase exact match	Precision	51.0±12.5% ***	0% ***
	Recall	43.8±11.2% ***	5.3±5.3% ***
	Number of return	1.1±0.5	0.05±.05
mean relevant return		2.7 ± 0.7	1.2±0.2

** p<0.01, *** p<0.001comparing phrase-based to the word-based matching method.
[#]: 2-component query has a pre-modifier(s) + head-noun structure (e.g. postsynaptic inhibitory synapse);
[##]: 3-component query has a pre-modifier + head-noun + post-modifier structure (e.g. ligand binding in the frontal cortex).